**Capstone Project 1**

**Predicting Loan Default**

Contents

[Introduction 2](#_Toc3665723)

[The Lending Club Data Set 2](#_Toc3665724)

[Problem Statement 3](#_Toc3665725)

[Assumptions 3](#_Toc3665726)

[Data Wrangling 4](#_Toc3665727)

[Exploratory Data Analysis 4](#_Toc3665728)

[Data Modelling 8](#_Toc3665729)

[Data Modelling Results 9](#_Toc3665730)

[1. Area Under ROC Curve 9](#_Toc3665731)

[2. Root Mean Square Error 10](#_Toc3665732)

[3. Accuracy Score 11](#_Toc3665733)

[4. Precision and Recall 12](#_Toc3665734)

[5. Confusion Matrix 13](#_Toc3665735)

[Conclusion 15](#_Toc3665736)

[References 15](#_Toc3665737)

# Introduction

The Lending Club is the world's largest peer-to-peer lending platform. Lending Club enables borrowers to create unsecured personal loans between $1,000 and $40,000. The standard loan period is three years. Lending Club enables borrowers to create loan listings on its website by supplying details about themselves and the loans that they would like to request. On the basis of the borrower’s credit score, credit history, desired loan amount and the borrower’s debt-to-income ratio, Lending Club determines whether the borrower is credit worthy and assigns to its approved loans a credit grade that determines payable interest rate and fees. The standard loan period is three years; a five-year period is available at a higher interest rate and additional fees. The loans can be repaid at any time without penalty.

Investors make money from interest. Rates vary from 6.03% to 26.06%, depending on the credit grade assigned to the loan request. Lending Club makes money by charging borrowers an origination fee and investors a service fee. The size of the origination fee depends on the credit grade and ranges to be 1.1%-5.0% of the loan amount. The size of the service fee is 1% on all amounts the borrower pays.

Loan delinquency is a failure to make loan payments when they are due. Extended delinquency can result in loan default. A lender may take legal action to get the money back. Load default has several adverse effects.

To reduce loan default risk, Lending Club focuses on high-credit-worthy borrowers, declining the loan applications, assigning higher interest rates to riskier borrowers within its credit criteria. Only borrowers with FICO score of 660 or higher can be approved for loans. The statistics on Lending Club's website state that, as of December 31, 2016, 62.3 percent of borrowers report using their loans to refinance other loans or pay credit card debt.

# The Lending Club Data Set

The Dataset used in this project is The Lending Club dataset available from kaggel website. (https://www.kaggle.com/wordsforthewise/lending-club/home). It is a real world data set which contains 2004126 rows of loan listing and 150 columns (attributes) of the each loan listing from year 2007 to 2018 Q2. Out of 150 columns, some columns have missing data and some columns are not needed for the analysis. After applying several data wrangling method, data set contains 2004095 rows and 83 columns. 4 columns are of data type *category*, 4 columns are of data type *datetime64*, 66 are of data type *float64*, and 9 are of data type *object*. Some of the interesting columns are:

|  |  |
| --- | --- |
| Id | A unique LC assigned ID for the loan listing. |
| loan\_amnt | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| funded\_amnt | The total amount committed to that loan at that point in time. |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| int\_rate | Interest Rate on the loan |
| grade | LC assigned loan grade |
| emp\_length | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. |
| home\_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| loan\_status | Current status of the loan |
| dti | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income. |
| delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| earliest\_cr\_line | The date the borrower's earliest reported credit line was opened |
| fico\_range\_low | The lower boundary range the borrower’s FICO at loan origination belongs to. |
| inq\_last\_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| mths\_since\_last\_delinq | The number of months since the borrower's last delinquency. |
| revol\_bal | Total credit revolving balance |
| total\_acc | The total number of credit lines currently in the borrower's credit file |
| out\_prncp | Remaining outstanding principal for total amount funded |

# Problem Statement

The Lending Club operates an online lending platform that enables borrowers to obtain a loan and investors to purchase notes backed by the payments made on loans. Investors can decide on investing on a loan applications based on the credit history, FICO score, income, current job status and many other attributes of borrowers mentioned in the application. This process comes with high risk of borrowers defaulting loans. Hence there is a need to classify each borrower as defaulter or not using the data collected when loan application is submitted. The purpose of this exercise is to predict if borrower will default the loan or not and help investors make a decision based on that.

# Assumptions

The Lending Club Dataset of all accepted loans from year 2007 to Q2 2018 is an imbalance dataset with very small number of loan applications which were defaulted. So we used the 'Charged off' loan status as an indicator of default.

Most of the features in the dataset represented information which are not available to investors at the time of loan application. These features were provided by The Lending Club after or during the loan approval process. So we used the features 'fico score' and 'dti' (Debt to Income Ratio) which are available to investors as the independent variables for our analysis.

# Data Wrangling

The Lending Club dataset is a real world data set and is relatively large data set. Lots of columns have missing values. Also some of the columns are not relevant for the analysis. So during data wrangling process, some columns and rows are dropped to get rid of missing values and non-relevant features.

* Converted columns such as ‘loan\_status’, ‘application\_type’ into categorical data type because Categorical data takes less memory and processing speed is faster.
* Converted columns representing dates such as ‘issue\_d’, ‘earliest\_cr\_line’ into datetime64 format to make date/time related processing easier.
* Filled NAN values with mean values of corresponding columns.
* Redundant variables such as loan grades and loan sub grades are dropped.
* The Pearson correlation between ‘fico\_range\_low' and 'fico\_range\_high' shows that both are correlated. So combined two features which is average of these features.
* The response variable in our dataset is ‘**loan\_status**’ which shows the status of the loan. It has 4 different values – ‘Charged Off’, ‘Fully Paid’, ‘Current’ and ‘Default’. Since our project goal is to predict whether a borrower will default the loan, we have changed ‘Fully Paid’ and ‘Current’ as 0 and ‘Charged Off’ and ‘Default’ as 1 where 1 indicates the borrower as a defaulter.

# Exploratory Data Analysis

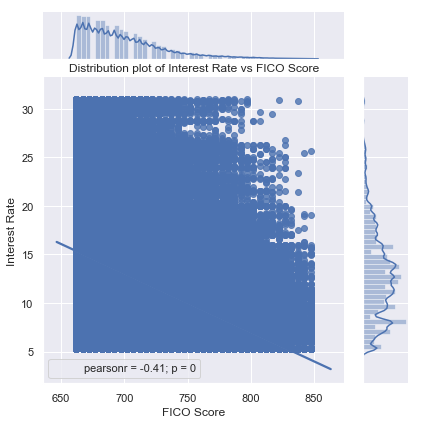
Exploratory data analysis was first defined by John Tukey in 1961 as *"Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."*

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. We can say that EDA is statisticians’ way of *storytelling* where you explore data, find patterns and tells insights. Often you have some questions in hand you try to validate those questions by performing EDA.

Initial exploration of the data revealed many interesting trends and findings.

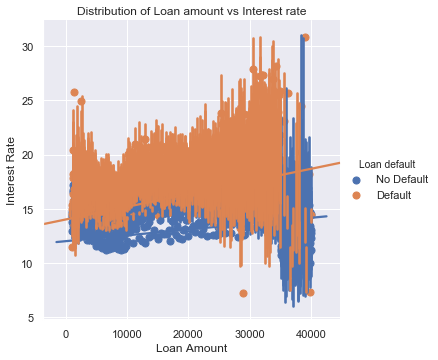
1. Distribution plot of Interest Rate vs FICO Score

*This plot shows that borrowers with lower FICO Score got high interest rate. The plot shows the distribution of FICO score on marginal x axis and distribution of Interest rate on marginal y axis. Pearson coefficient is -ve which means increasing FICO score corresponds to lower interest rate.*



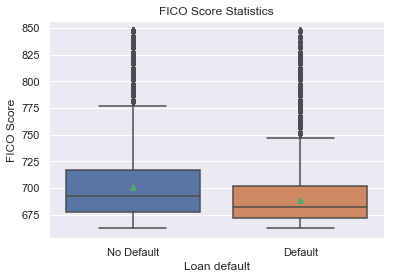
1. Distribution of Loan amount vs Interest rate for default and no default loans

*The plot shows that the same range of loan amounts have higher interest rate for loans which are defaulted.*

******

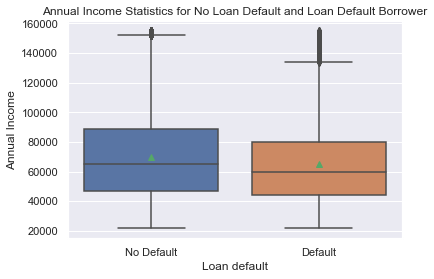
1. FICO Score Statistics for default and No Default loans

*The whisker (box) plot of FICO Score displays, minimum, maximum, median, 1st quantile, 3rd quantile values. Since mean and median value of loans is less for loans which are defaulted, the plots show that borrowers with less FICO score have higher chances defaulting the loan.*



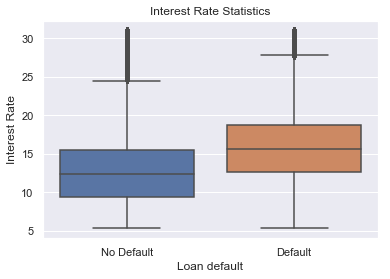
1. The whisker plot for Annual Income Statistics

*The plot shows that the mean and median of annual income of borrowers with defaulted loan are lower than the mean and median of the annual income of borrowers with no default. But this difference is not significantly big that we can draw a clean conclusion based on it.*



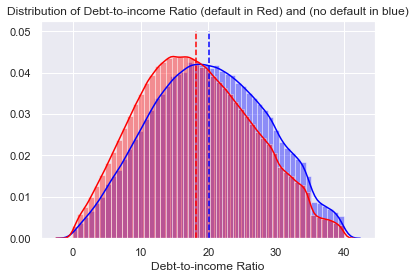
1. The whisker plot for Interest Rate Statistics.

*The plot shows that minimum and maximum values of interest rates are similar but the mean and median of interest rates of loans with default are significantly higher than the mean and median of the interest rates of loans with no default.*



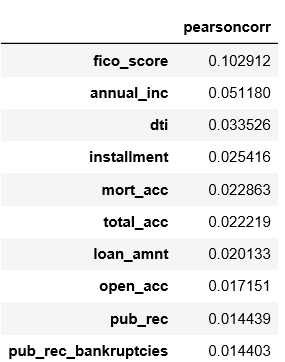
1. The distribution graph for debt to interest ratio for default and not default borrowers

*The plot shows the distribution graph is normal for both default and no default borrowers. The default borrowers tend to have higher dti ratio than the no default borrowers.*



1. Findings and Conclusion

*The Pearson Correlation Coefficient shows that the features such as fico\_score, annual\_inc and dti (Debt to Income Ratio) are the most interesting features which are positively related to default loan status.*



# Data Modelling

* As we found out from EDA, fico\_score and dti are two valuable features that we are going to use to model and predict if loan is going to default or not.
* The data set has very small number of default loans and that’s not enough to train the classification models. There are significant amount of loans which are categorized as charged off. We used the charged off loan status as default status and trained the model and verified.
* This is classification problem with two possible target values of 0 = no default and 1 = default. We will be using several classification algorithms from scikit-learn library.
* Also this is an imbalanced dataset with less than 10% rows with default = 1 values. So we take a subset of the loan data which has equal number of default and no default rows.
* This is a classification problem of predicting if a loan will default or not. So different models are build, fit and validated. To do the cross validation, we split the data set in training and test data sets in 80/20 ratio.
* Simple Classification algorithms such as Logistic Regression and K Nearest Neighbors are used to model the data. Complex Ensemble classification algorithms such as Random Forest, Bagging Classifier with KNN, AdaBoost with DecisionTree are also modeled to improve the performance of algorithms.
* GridSearchCV is used to tune hyper-parameters with cross validation at the same time.

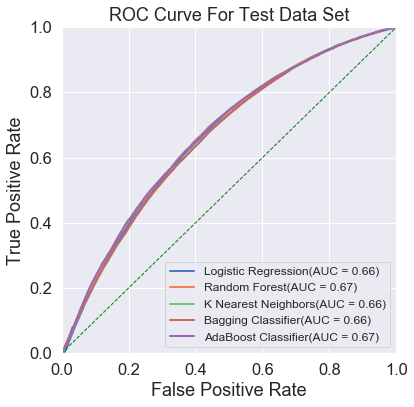
# Data Modelling Results

To evaluate the performance of models, several metrics are measured. For the Loan default prediction, Recall i.e.False Negatives Rate is the best metric to evaluate the model. Lower the number of false negatives, better the model is. Here, False negative is when model predicting “a borrower will not default a loan even though he will “. Our model cannot afford having higher False Negatives as it leads to negative impact on the investors. So, we evaluated our models using the number of False Negatives and accuracies and other metrics.

## Area Under ROC Curve

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The best possible prediction method would yield a point in the upper left corner or coordinate (0, 1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The diagonal divides the ROC space. Points above the diagonal represent good classification results; points below the line represent bad results. A predictor with largest AUC value performs the best.

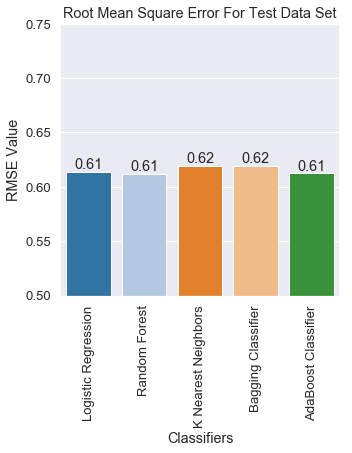
In the plot below, AdaBoost Classifier and Random Forest classifiers performed best with AUC = 0.67. Logistic Regression, K Nearest Neighbors and Bagging Classifier algorithms performed second best with AUC = 0.66



## Root Mean Square Error

RMSE of an estimator measures the square root of average of the squares of the errors—that is, the square root of average squared difference between the estimated values and actual values. The RMSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.

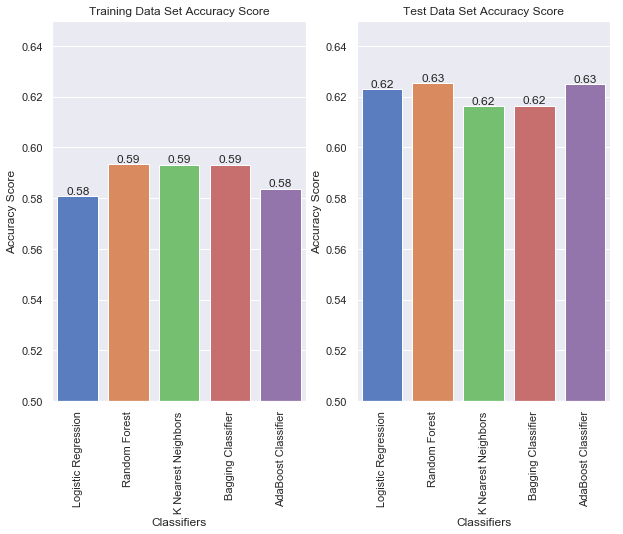
In the bar graph below AdaBoost, Random Forest and Logistic Regression performed best as minimum RMSE value of 0.61.



## Accuracy Score

Classification accuracy score is the number of correct predictions made divided by the total number of predictions made.

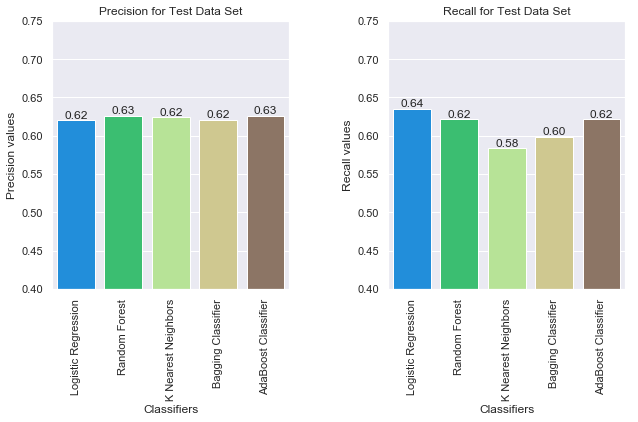
The bar graphs below show the comparison of accuracy scores for both training data set and test dataset. AdaBoost, and Random Forest classifiers performed best with accuracy score of 0.63 on test data set.



## Precision and Recall

Precision is about how precise/accurate the model is out of those predicted positive, how many of them are actual positive. Recall actually calculates how many of the Actual Positives the model capture through labeling it as Positive (True Positive). Recall shall be the model metric used to select the best model when there is a high cost associated with False Negative.

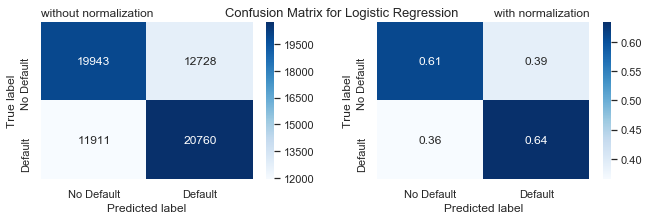
The comparison of Precision and Recall values below in the bar graphs show that Logistic Regression performed best with Recall value of 0.64.

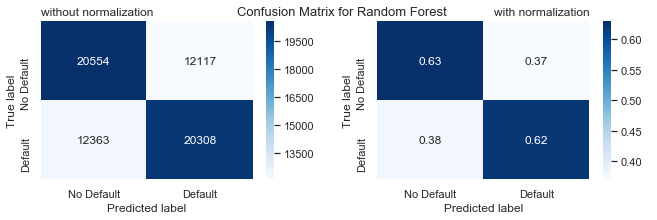


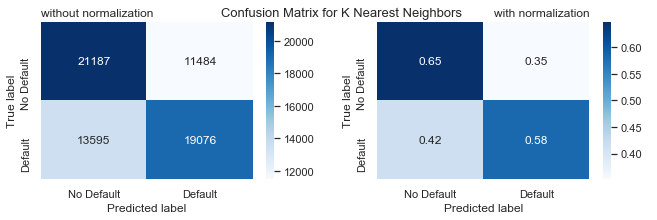
## Confusion Matrix

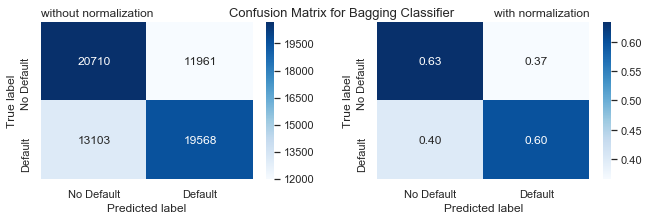
A table of confusion or confusion matrix, is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives. This allows more detailed analysis than mere proportion of correct classifications (accuracy).

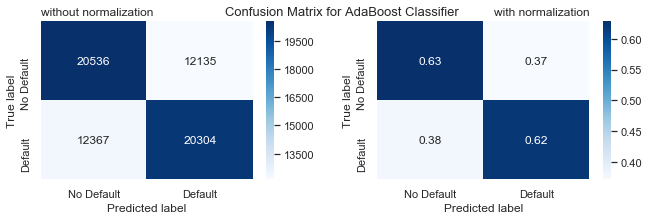
The plots below show the confusion matrix for each of the predictors.











# Conclusion

###### Predicting Loan Default in The Lending Club dataset is a real business problem. The Loan entries in the dataset have already been screened through the existing algorithms to predict the loan defaults which makes 'predicting loan default' from this dataset a highly difficult problem. This exercise tried to estimate an improvement over the existing solution.

###### I ran 5 classification algorithms and got best AUC score of 0.67, best accuracy score of 0.63, best recall value of 0.64 and best precision value of 0.63. These scores are good considering the fact that the problem is highly difficult real business problem.

###### The above graph shows that among all the loans accepted in first 2 quarters of 2018, 202 loans for a total of almost 3 million dollars are predicted default correctly. During the year 2017, 2864 loans for a total of almost 42 million dollars are predicted default correctly. So this exercise actually add values to The Lending Club business.

# References

1. <https://www.lendingclub.com/>
2. <https://en.wikipedia.org/wiki/Lending_Club>
3. [https://scikit-learn.org/stable/index.html#](https://scikit-learn.org/stable/index.html)
4. <https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html>
5. <https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>
6. <http://seaborn.pydata.org/index.html>